Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy

Year: 2023 | *Citations: 366* | *Authors: Zhihong Shao, Yeyun Gong, Yelong Shen*

Abstract

Large language models are powerful text processors and reasoners, but are still subject to limitations including outdated knowledge and hallucinations, which necessitates connecting them to the world. Retrieval-augmented large language models have raised extensive attention for grounding model generation on external knowledge. However, retrievers struggle to capture relevance, especially for queries with complex information needs. Recent work has proposed to improve relevance modeling by having large language models actively involved in retrieval, i.e., to improve retrieval with generation. In this paper, we show that strong performance can be achieved by a method we call Iter-RetGen, which synergizes retrieval and generation in an iterative manner. A model output shows what might be needed to finish a task, and thus provides an informative context for retrieving more relevant knowledge which in turn helps generate a better output in the next iteration. Compared with recent work which interleaves retrieval with generation when producing an output, Iter-RetGen processes all retrieved knowledge as a whole and largely preserves the flexibility in generation without structural constraints. We evaluate Iter-RetGen on multi-hop question answering, fact verification, and commonsense reasoning, and show that it can flexibly leverage parametric knowledge and non-parametric knowledge, and is superior to or competitive with state-of-the-art retrieval-augmented baselines while causing fewer overheads of retrieval and generation. We can further improve performance via generation-augmented retrieval adaptation.